



IN4325

Information Retrieval

Claudia Hauff (WIS, TU Delft)

- IR deals with the representation, storage, organization of, and access to largely **unstructured** information
- Central notions are **information needs** and **relevance**
- IR has its roots in the **library and information sciences**



About 10.200 results (0,31 seconds)

Did you mean: **time slot booking** not doodle

Online appointment scheduling - Doodle

<https://doodle.com/free-online-appointment-scheduling> ▼

No more confusion, no more missed appointments. ... and booked **time slots** on a piece of paper is not conducive for amendments and changes of plans.

People also ask

How do you schedule a doodle poll?

How do I schedule an appointment?

Why is appointment scheduling important?

How do I make a booking website?

Images for timeslot booker not doodle



➔ More images for timeslot booker not doodle

The Doodle Web Scheduler

<https://doodle.com/web-scheduler> ▼

No more missed appointments, no more cluttered diaries. ... system will allow you to use the function that allows participants to only choose one **time slot**. Have a ...

Missing: ~~booker~~ | Must include: **booker**

crawling and indexing

vertical selection

query suggestions

result ranking

snippet generation

implicit feedback

Report images

entity cards

About 9.150.000 results (0,48 seconds)

Online appointment scheduling - Doodle

<https://doodle.com/free-online-appointment-scheduling> ▼

No more confusion, no more missed appointments. ... and booked time not conducive for amendments and changes of plans.

quick and easy online booking system - Doodle

<https://doodle.com/online-booking-system> ▼

Use Doodle to schedule events with friends and colleagues. ... Doodle where you can quickly and without hassle arrange a time to ... most fun fun and relaxed!) online booking system on ...

Staff scheduling made easier with Doodle

<https://doodle.com/staff-scheduling> ▼

Doodle already has a staff schedule template ready for you to adapt and be confident that your staff scheduling will be done on time and in a ... No more and no more double-fudge sundae for you!

Free Online Booking Software from Doodle

<https://doodle.com/booking-software> ▼

Doodle Booking Software lets you easily find time to meet up with friends your online calendar and schedule your week all in one place. ... and 'n marking their availability for each time.

Easily arrange meetings with Doodle's appointment

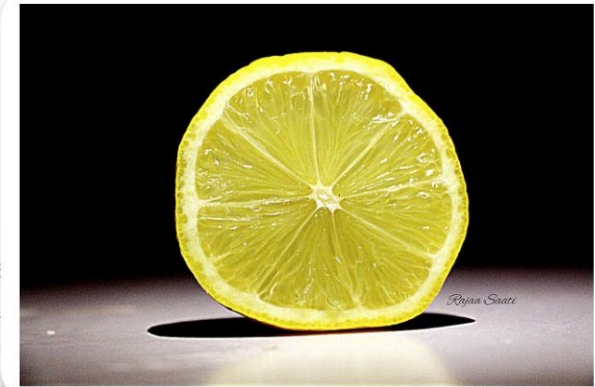
<https://doodle.com/appointment-calendar> ▼

Arrange meetings and organise your schedule with Doodle's simple ... when colleagues or friends are not immediately ... Once created, you can colleagues and decide upon which time is best ...

Organise your work schedule with Doodle

<https://doodle.com/work-schedule> ▼

On Doodle.com you can set up an event poll with various dates and times your office team by showing when you are not available ... schedule so



Please wait for a few seconds
generated for the user

Caption: a close up of a red and white flower

Is it a flower?

yes

What is yellow in the image?

it's just a close up of the bird

Which bird?

i ca n't tell

core IR
module

applied NLP
module

Course setup

			<i>grading</i>
			<i>45%</i>
W3.1-3.4 core IR	group (2-3) project	paper reviews	
Claudia Hauff			<i>10%</i>
W3.5-3.8 applied NLP	group (3-4) project	paper reviews	
Nava Tintarev			<i>45%</i>

To pass: grade of **5+** in both group projects and **9** out of 14 offered paper reviews completed with a *sufficient*.

Weekly support hours, starting in week 3.2.

Core IR project

detailed

There are three expected outputs:

instructions

1. **Project proposal** (mandatory, but ungraded - you will receive feedback)
2. **Intermediate project report** (mandatory, but ungraded - you will receive feedback)
3. **Final project report** (mandatory, graded).

Group projects are conducted in groups of 2-3 students. Please enroll to Brightspace - we have predefined 40 different groups, take your pick!

You can choose between two types of projects: reproducing an IR paper or your own.

Restrictions

Based on prior experience we put the following restrictions on your choice of project to follow your own research idea:

- You can only conduct a project on neural IR if you have successfully completed the neural IR course
- The main focus of your project is *IR* (testing different retrieval models, evaluation, user interfaces, exploring different modes of searching, etc.). Recommender systems are not allowed
- The project makes use of an **actual search engine** (i.e. you set up your own search engine or use a sufficient to retrieve search results from the Bing API).
- The project focuses on **textual data** (not video/audio/genomics/.... data)

We have a first deadline in week 3.2, so you will get early feedback on the success of your project

Table of contents

- IR resources
 - Books
 - Software
 - Datasets

*everything
online*

- Lecture 1: evaluation (week 3.1)
 - Recommended readings
- Lecture 2: classic retrieval models (week 3.1)
 - Recommended readings
 - ⚠ Paper P1 to review
- Lecture 3: indexing (week 3.2)
 - Recommended readings
 - ⚠ Paper P2 to review
- Lecture 4: query refinement (week 3.2)
 - Recommended readings
 - ⚠ Paper P3 to review
- Lecture 5: interactive IR (week 3.3)
 - Recommended readings
 - ⚠ Paper P4 to review
- Lecture 6: personalization (week 3.3)
 - Recommended readings
 - ⚠ Paper P5 to review
- Lecture 7: learning to rank (week 3.4)
 - Recommended readings
 - ⚠ Paper P6 to review
- Lecture 8: neural IR (week 3.4)
 - Recommended readings
 - ⚠ Paper P7 to review



Brightspace

- Forming project groups
- Submission of reviews / project reports
- Q&A forum (we prefer Slack)
- Grading

Slack: in43252019.slack.com

- Questions to the course team

Email: in4325-ewi@tudelft.nl

- Responsible instructors only



IN4325

Evaluation in IR

Claudia Hauff (WIS, TU Delft)

The big picture

The essence of classic IR

Information need: *Looks like I need Eclipse for this job. Where can I download the latest beta version for macOS Sierra?*

Information need

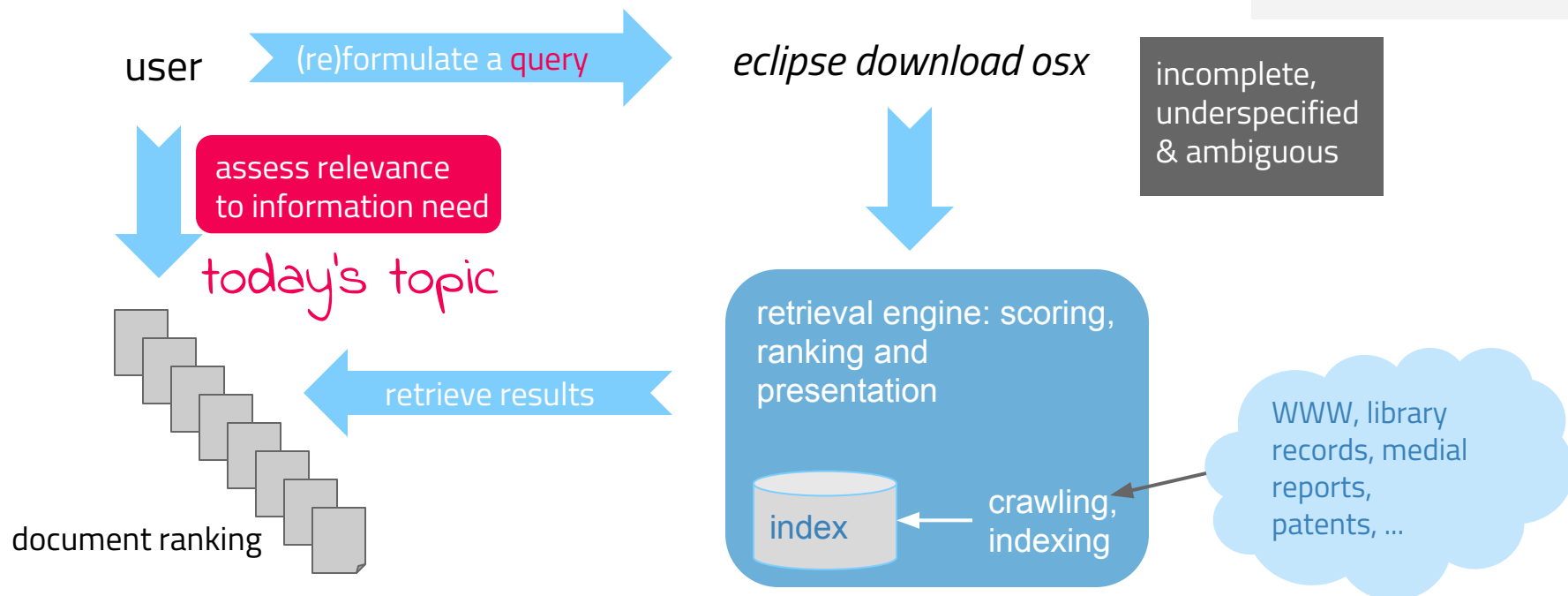
Topic the user wants to know more about

Query

Translation of need into an input for the search engine

Relevance

A document is relevant if it (partially) provides answers to the information need



Why “classic”?

Classic Web search	Proactive search (zero query search)	Voice search
<u>Query</u> = textual input	<u>Query</u> = none	<u>Query</u> = speech input
<u>Results</u> = ranked list of search result snippets (i.e. “ten blue links”)	<u>Results</u> = a single information card	<u>Results</u> = speech output
<u>Actions</u> = click, view	<u>Actions</u> = view	<u>Actions</u> = speech input

...

Information retrieval is a broad field that deals with a wide range of information access issues.

Connected to information science, NLP, applied machine learning, semantic Web and (in recent years) dialogue systems.

What are we up to as IR community?

Let's quickly look at upcoming benchmark tasks (TREC* 2018)

Complex Answer Retrieval

“ The focus ... is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus. ”

Incident Streams

“ ... to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests ... for emergency service operators. ”

Precision Medicine

“ ... building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments ”

News search

“ ... will foster research that establishes a new sense what relevance means for news search. ”

* Text REtrieval Conference (1992 - *), changing tracks every year. trec.nist.gov

Benchmarks drive our community

CLEF

Conference and Labs
of the Evaluation
Forum

<http://www.clef-initiative.eu/>

EUROPE

MediaEval

Benchmarking
Initiative for
Multimedia
Evaluation

<http://www.multimediaeval.org/>

EUROPE

NTCIR

NII Test Collection for
IR Systems

<http://research.nii.ac.jp/ntcir/index-en.html>

JAPAN

FIRE

Forum for Information
Retrieval Evaluation

<http://www.isical.ac.in/~cia/>

INDIA

TREC

USA

TRECVID

USA



*' ... engineers then come up with a **hypothesis** about what signal what data could we integrate into our algorithm we test all these reasonable ideas through **rigorous scientific testing** ... ' Google Inside Search*

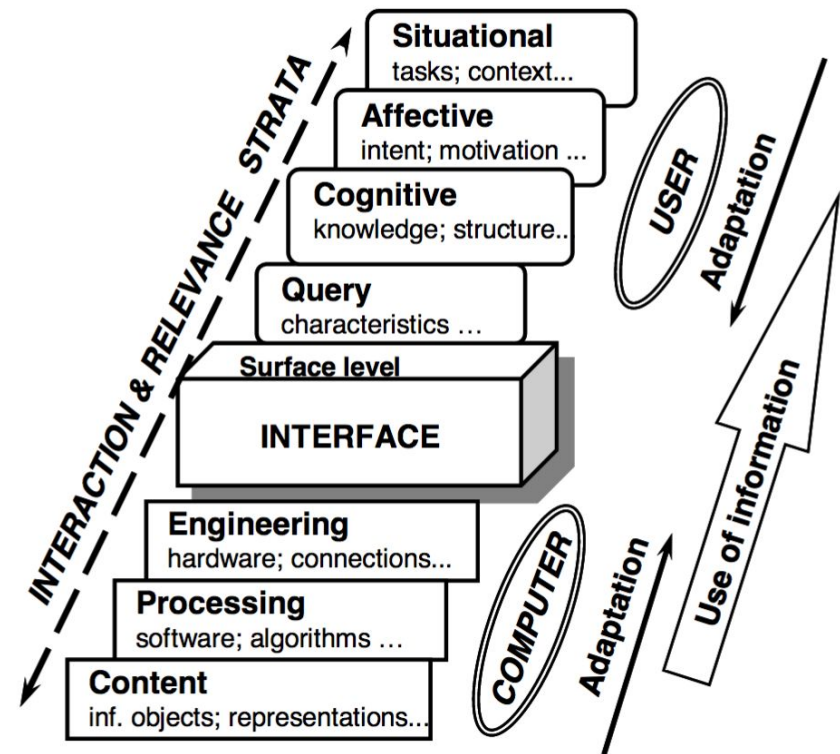
Relevance

Why are we starting with the evaluation lecture in this course anyway?

Because evaluation is a vital component of ~95% of all published IR papers. No matter your choice of project or survey, you need to understand IR evaluation.

Relevance

- Key notion in information retrieval
- A good retrieval system retrieves all relevant documents but as few non-relevant documents as possible
- Relevance is an intuitive notion for humans
- Retrieval systems **create** relevance and users **derive** relevance
- Ongoing debate for the past 40 years



Stratified model of relevance interactions.
(Saracevic, 2007)

Manifestations of relevance (Saracevic, 2007)

- **System relevance:** relation between query and information objects (documents)
- **Topical relevance:** relation between the subject of the topic and the subject of the information objects
- **Cognitive relevance:** relation between the cognitive state of the user and the information objects
→ cognitive correspondence, novelty, information quality, ...
- **Situational relevance** (utility): relation between the situation and the information objects
→ appropriateness of information, reduction of uncertainty, ...
- **Affective relevance:** relation between the intent, goals, emotions of the user and information
→ success, accomplishment, ...

Evaluation is at the heart of IR

Goals

Evaluation measures that reflect users' satisfaction with the system

The perfect metric also allows us to fine-tune the system via machine learning

User satisfaction in terms of

- **Coverage** of the corpus
 - **Time lag** between query and retrieved results - even 200ms delays are noticeable to users ⁽¹⁾
 - **Presentation** of output
 - Required **user effort**
 - ...
 - Proportion of relevant results retrieved (**recall**)
 - Proportion of retrieved results that is relevant (**precision**)
 - ...
- system effectiveness*

Assumption: the more effective the system, the more satisfied the user.

Evaluation is difficult

- Which users to evaluate for?
- Which intents to evaluate for?
- How are evaluations be made reusable?
- How can the difference between systems be quantified?

Test Collection Approach *

* Mainstream way of evaluation. Empirical. Another approach is the **axiomatic one** (found in theoretic research).

Cranfield evaluation paradigm (1960s)

IR evaluation methodology developed by Cyril Cleverdon in the 1960s; Cranfield corpus:

- Test collection of 1,400 documents (1) [scientific abstracts]
- Set of 225 topics (information needs)
- **Ad hoc** task
- **Complete** set of binary (0/1) relevance judgments
- Metrics to compare systems with each other
- **i.e. reusable data!**

Example Cranfield corpus topic:

papers applicable to this problem (calculation procedures for laminar incompressible flow with arbitrary pressure gradient)



flickr@mkgrimaldos

Paradigm adapted to the modern time

Relevance judgments are
no longer binary

- **Multi-graded** decision
(somewhat relevant vs. very relevant)
- **User-dependent** decision (what is relevant for you may not be relevant for me)
- **Context-dependent** decision
(whether something is relevant depends on the time of day, ...)

Topics and queries are not one and
the same anymore

TREC 2001 Web ad hoc topic

<top>

<num> Number: 503

<title> Vikings in
Scotland?

<desc> Description:
What hard evidence
proves that the Vikings
visited or lived in
Scotland?

<narr> Narrative: A document that merely states that the Vikings visited or lived in Scotland is not relevant. A relevant document must mention the source of the information, such as relics, sagas, runes or other records from those times.

</top>



We are conducting **simulations** of users searching with a retrieval system.

- + Cheaper, easier, reusable, reproducible
- Test collection retrieval effectiveness gains (=simple simulated users) may not translate to operational gains (=real users).

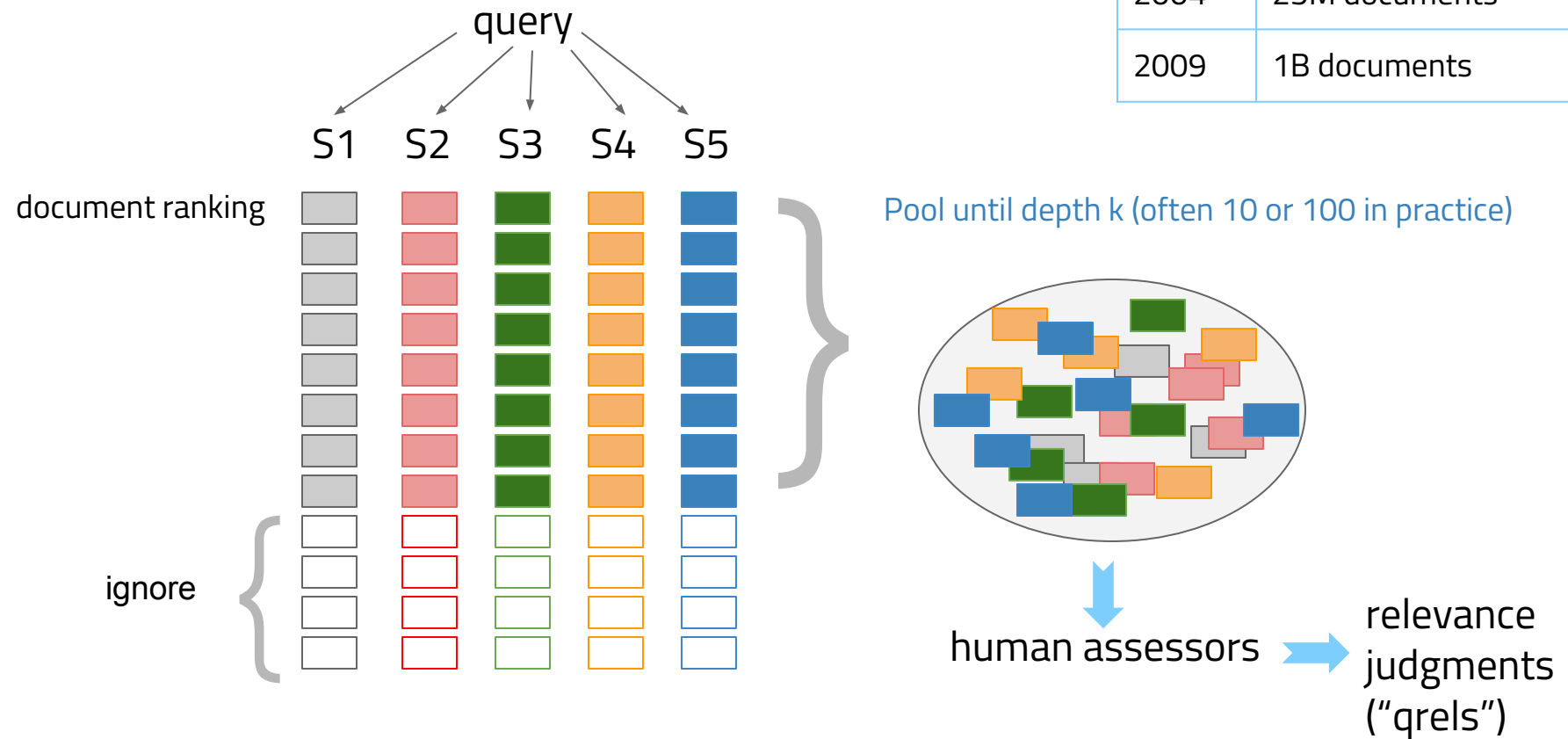
Also known as **batch evaluation** or **offline evaluation**.



What documents to judge: depth pooling

Commonly used today

Year	TREC Web corpus sizes
2001	1.69M documents
2004	25M documents
2009	1B documents



 50

Topics (ad hoc task)

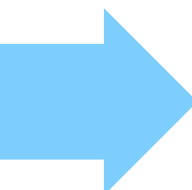
TREC-8 numbers
(ran in 1999)

 86,830

Pooled documents (k=100)

 129

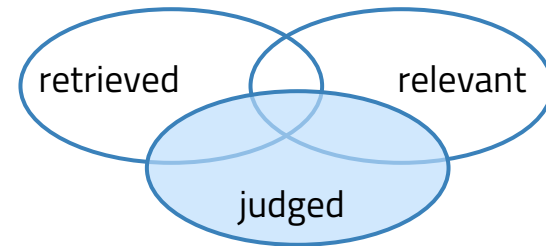
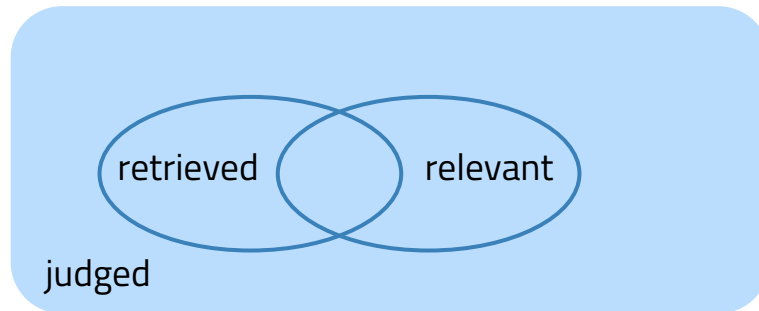
Systems

 723

Assessor hours

At \$20 an hour, that amounts to \$14,460.
And thus, we are still using the TREC-8
corpus to this day for experiments!

Cranfield vs. TREC depth pooling



Relevant documents not appearing in the pool are missed.

Test collections are vital to ensure continuous **algorithmic improvements** (one could argue) ...

however, papers are easier to publish when results are **positive**.

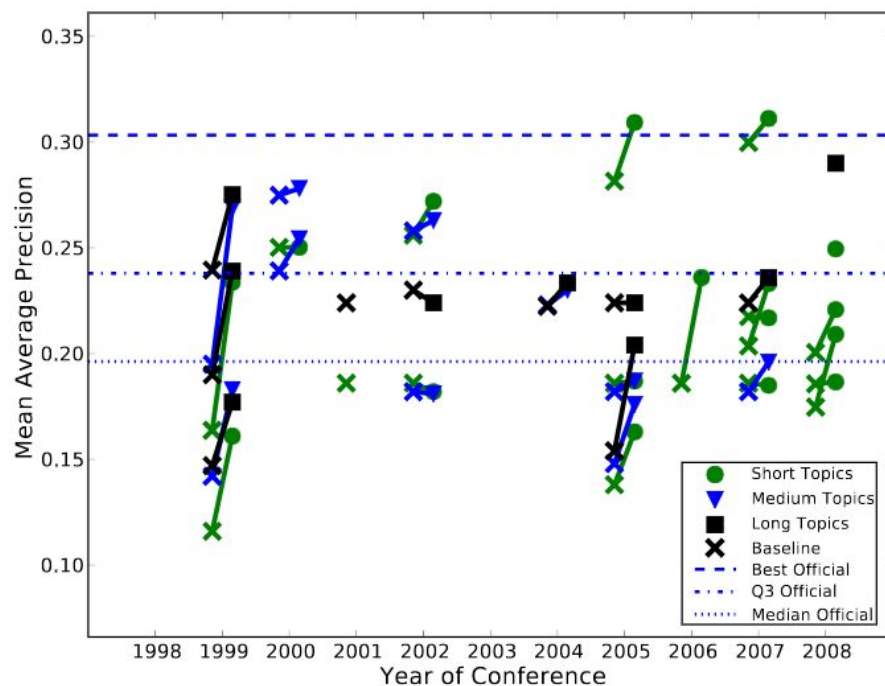


Figure 1: Published MAP scores for the TREC 7 Ad-Hoc collection. The connections show before-after pairs.



















































baseline vs. improved



Popular Evaluation Measures

There are 60+ published IR metrics, we picked 7 here do not despair by the number of metrics. It is a tiny part of what is out there.

Precision

	S1	S2	S3	S4	S5
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					

relevant
non-relevant

Precision at 10 docs
"P@10"

	S1	S2	S3	S4	S5
Precision at 10 docs	0.0	1.0	0.3	0.4	0.3

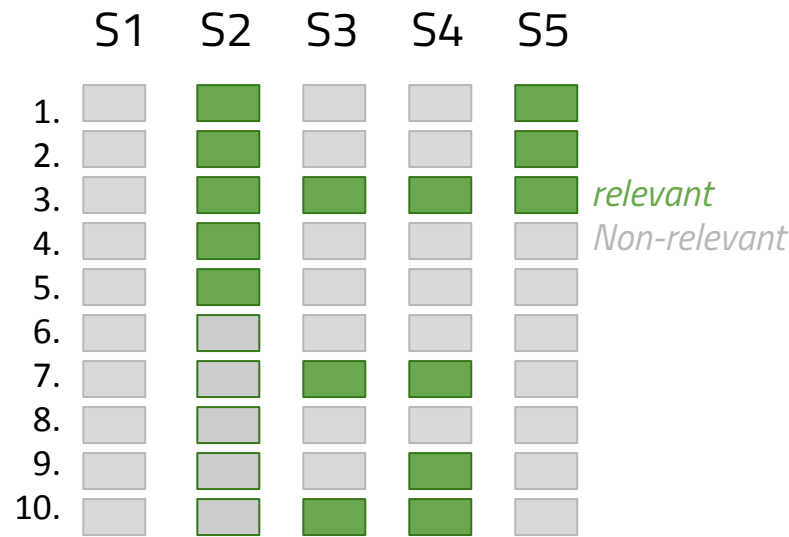
One query, five systems

Precision measures a system's ability to **only** retrieve relevant items.

R-precision is P@R where R=number of relevant documents.

$$\text{precision} = \frac{\text{num. relevant docs retrieved}}{\text{num. docs retrieved}}$$

Recall



Recall 0.0 1.0 0.6 0.8 0.6
(assume R=5)

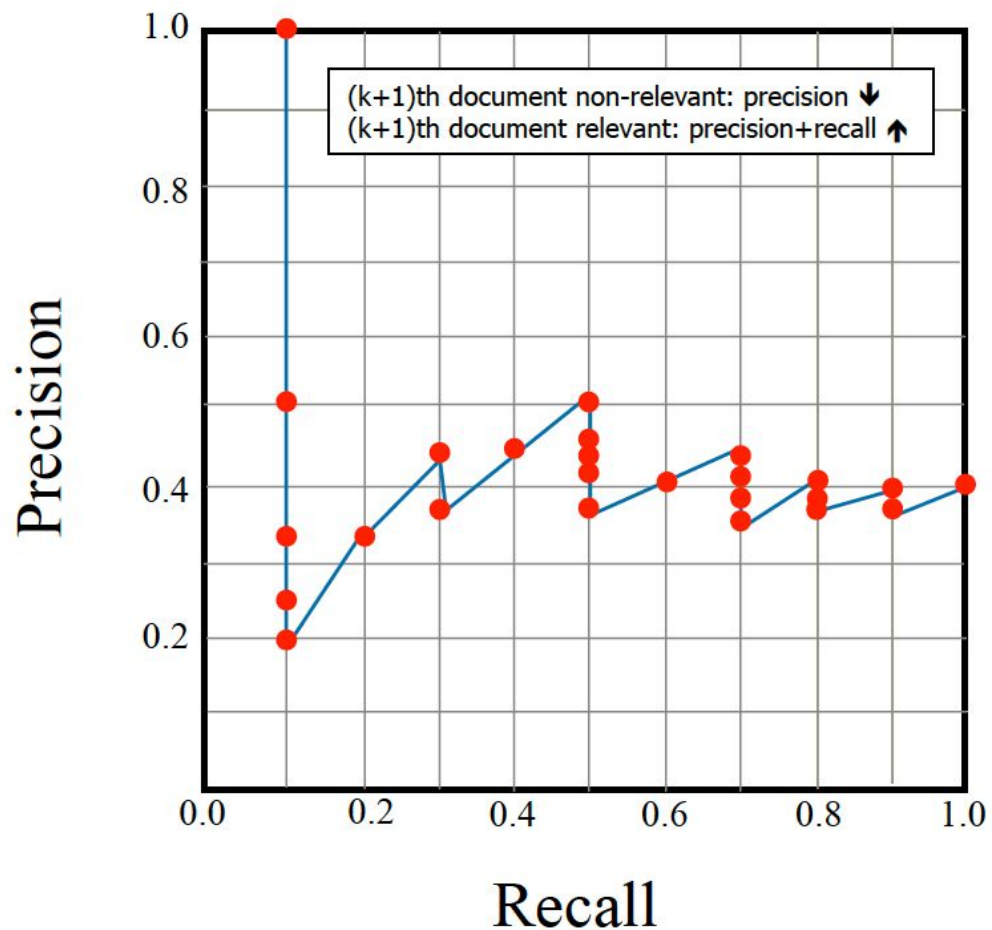
$$\text{recall} = \frac{\text{num. relevant docs retrieved}}{\text{num. relevant docs in corpus}}$$

One query, five systems

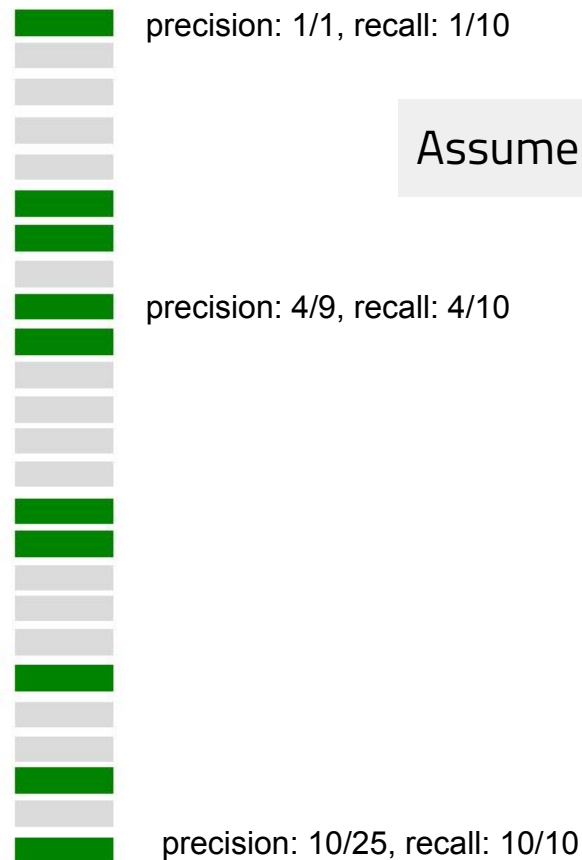
Recall measures a system's ability to retrieve all R relevant documents.

Recall and Precision are **set-based** measures. Retrieved are **ranked lists**.

Recall-Precision Curve

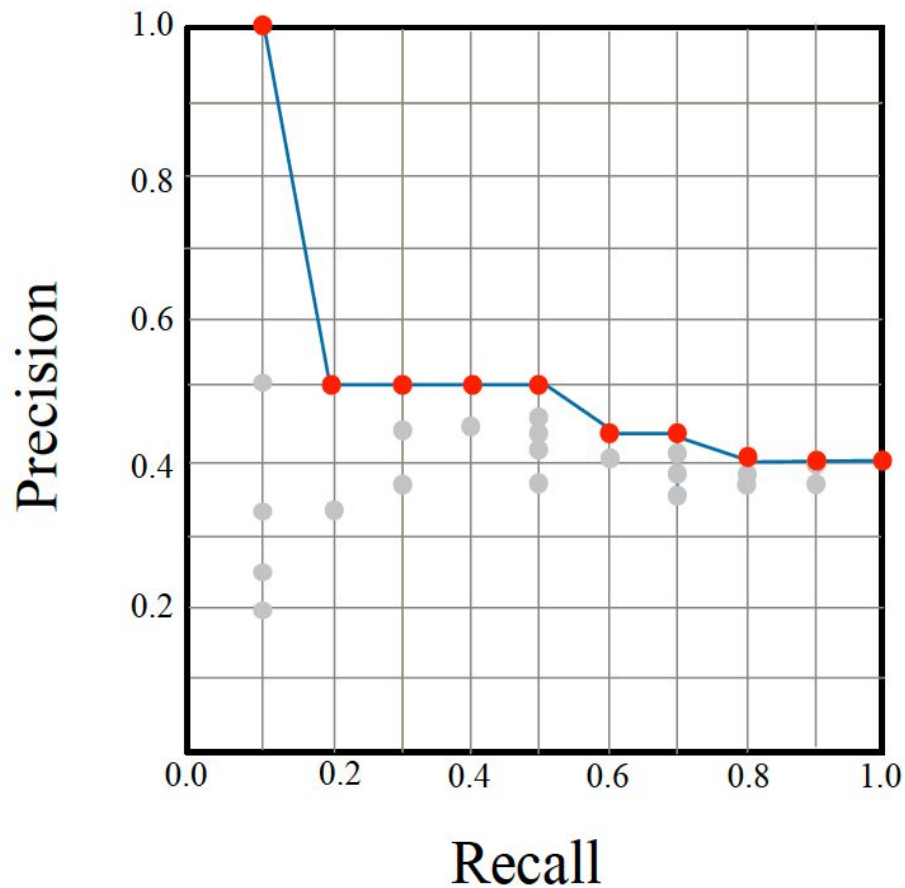


One query, one system



Recall-Precision Curve

One query, one system

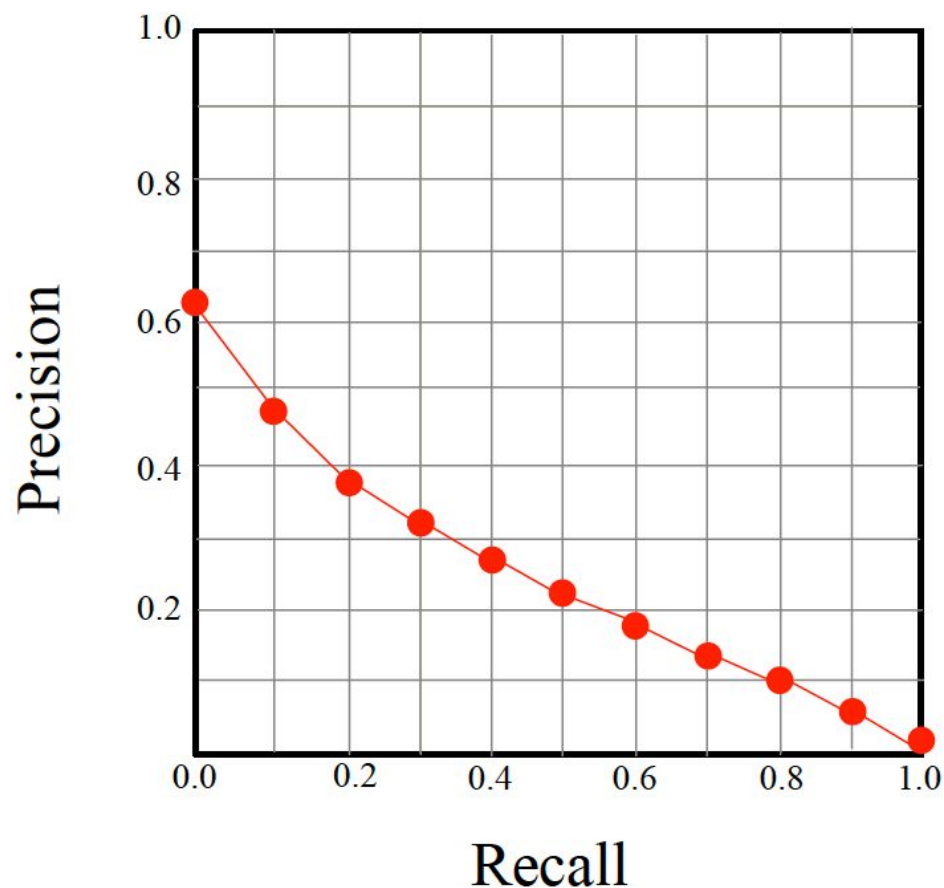


Interpolated precision at recall-level R

$$precision_{interp}(r) = \max_{r' \geq r} precision(r')$$

Recall-Precision Curve

Many queries, one system

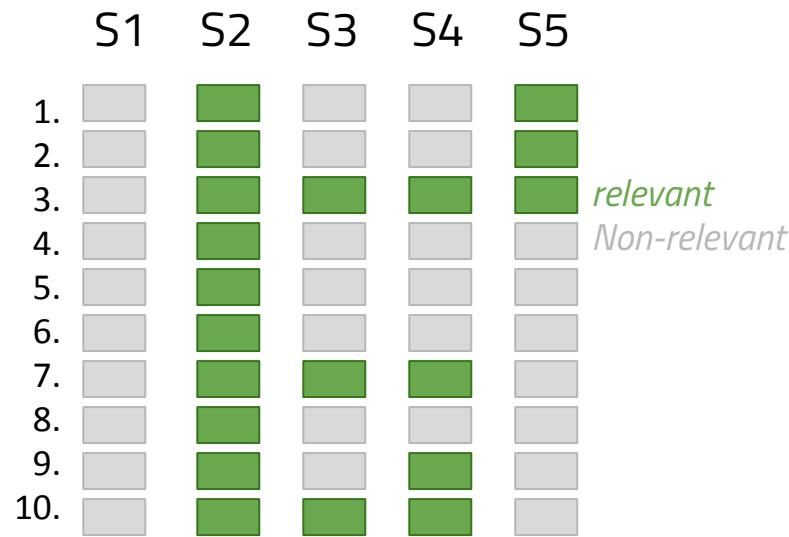


Precision at 11 standard recall values.
Averaged over all queries.

The more relevant documents are retrieved (recall \nearrow), the more non-relevant documents are retrieved (precision \searrow)

Problem: this is a graph, not a single number ... how do systems compare with different precision-recall curves?

Average Precision



AvP (assume R=10)

	S1	S2	S3	S4	S5
AvP	0.0	1.0	0.09	0.13	0.3

$$AvP = \frac{1/3 + 2/7 + 3/9 + 4/10}{10}$$

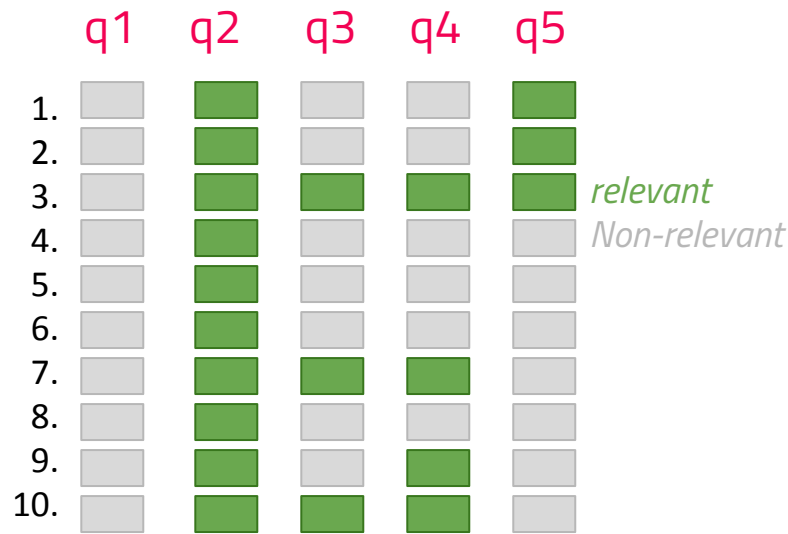
One query, five systems

Average precision takes the **order** (ranking) of the relevant and non-relevant documents into account

Average precision takes the **number R** of relevant documents into account.

$$AvP = \frac{\sum_{k=1}^s P@k \times rel(k)}{R}$$

Mean Average Precision



AvP 0.0 1.0 0.09 0.13 0.3
(assume R=10)

MAP = 0.364

One system, five queries

Given a **set of queries**, the average effectiveness is the **mean over AvP**.

MAP remains one of the most commonly employed retrieval evaluation measure to this day.

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{k=1}^s P@k \times rel(k)}{R}$$

Geometric Mean Average Precision

A measure designed to highlight improvements for low-performing topics

Geometric mean of per-topic average precision values (n is the num. topics):

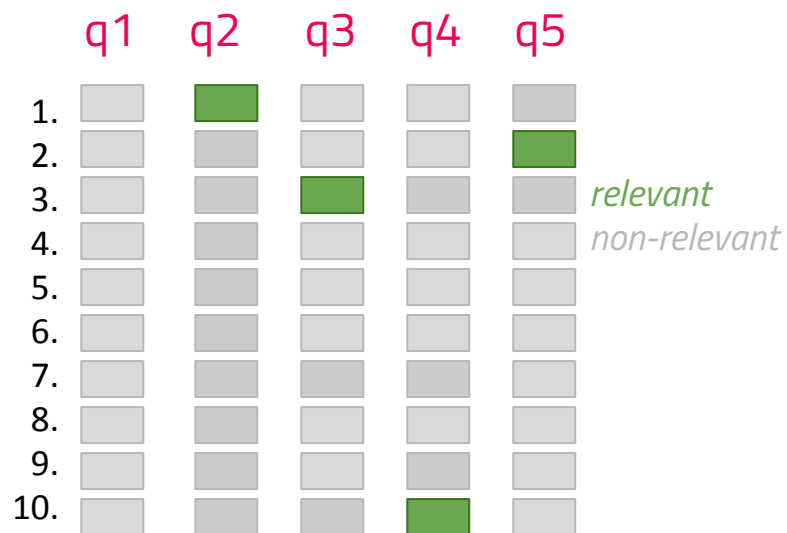
$$GMAP = \sqrt[n]{\prod_n AP_n}$$
$$= \exp \frac{1}{n} \sum_n \log AP_n$$

Two systems, five queries

	S1	S2	
q1	0.60	0.58	
q2	0.20	0.18	
q3	0.01	0.03	←
q4	0.04	0.06	←
q5	0.90	0.90	
MAP = 0.350		MAP = 0.350	
GMAP = 0.134		GMAP = 0.176	

S2 performs better on the **worst topics**! Can we have a measure that prefers systems that do well on the worst topics?

Mean Reciprocal Rank



RR 0.0 1.0 0.33 0.01 0.5

MRR=0.369

One system, five queries

One relevant document per query

Reciprocal rank averaged over all queries.

$$RR = \frac{1}{\text{rank of relevant doc}}$$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

- Standard Web search queries are short (2-3 terms), e.g. "cheap internet", "dinosaurs", "solar panels"
- **Graded** relevance scales needed (e.g. 0-3)
- NDCG measures the "gain" of documents
- Assumptions:
 - **Highly relevant** documents are more valuable than **marginally relevant** documents
 - The greater the ranked position of a relevant document, the less **valuable** it is for the user
 - Few users go further than the first 10 blue links
 - Probability of reaching the document is lower
 - Users have limited time
 - Users may have seen the information in the document already

Instead of just giving you the end-result, let's look at how the metric was **developed**.

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Direct cumulative gain can be defined iteratively



$$G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$$

$$CG_i = \begin{cases} G_i, & \text{if } i = 1 \\ CG_{i-1} + G_i, & \text{otherwise} \end{cases}$$

$$CG' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots \rangle$$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Discounted cumulative gain: reduce the document score as its rank increases (but not too steeply)

- Divide the document score by the log of its rank
- Base of the logarithm determines discount factor

$$DCG_i = \begin{cases} CG_i, & \text{if } i < b \\ CG_{i-1} + G_i / \log_b i, & \text{if } i \geq b \end{cases}$$

assume $b=2$

$CG' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots \rangle$

$DCG = \langle 3, 5, 6.9, 6.9, 7.3, 8, 8.7, 9.6, 9.6, \dots \rangle$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Normalized discounted cumulative gain: compare DCG to the theoretically best possible

- Ideal vector sorts the document relevance judgments in decreasing order of relevance

I' is based on the search topic,
not the retrieval result!

$I' = \langle 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, \dots \rangle$

$CG_{I'} = \langle 3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, \dots \rangle$

$DCG_{I'} = \langle 3, 6, 7.9, 8.9, 9.8, 10.5, 10.9, 11.2, 11.5, 11.8, \dots \rangle$

- The DCG vectors are divided component-wise by the corresponding ideal DCG vectors

Normalization so that a perfect
ranking at k for query j is 1

- NDCG for queries Q at rank k :

$$NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1+m)}$$

Relevance score assessors gave D at query j

Statistical significance tests

Significance tests

- Given the results from a number of queries, how can we conclude that ranking **algorithm A** is better than **algorithm B**?
- Significance tests enable us to reject the **null hypothesis** (no difference) in favor of the **alternative hypothesis** (B is better than A)
- (`trec_eval` does not come with those)

Significance tests

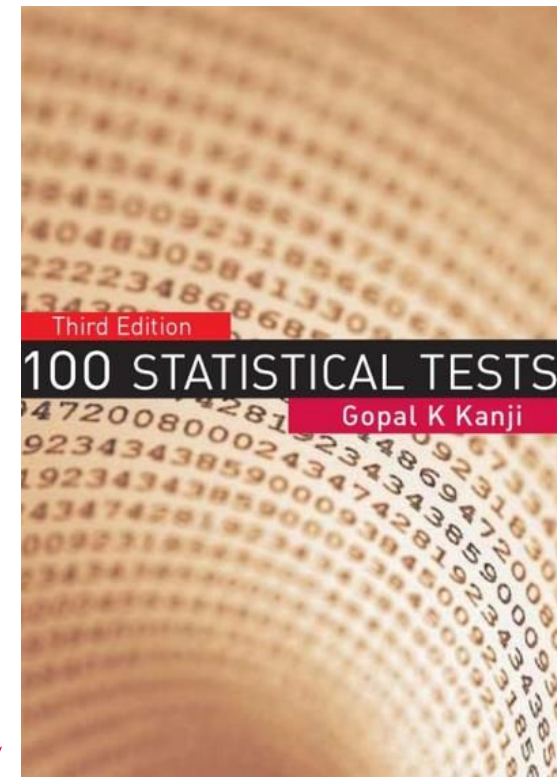
1. Compute the effectiveness measure for every query for both rankings.
2. Compute a *test statistic* based on a comparison of the effectiveness measures for each query. The test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide whether or not the null hypothesis should be rejected.
3. The test statistic is used to compute a *P-value*, which is the probability that a test statistic value at least that extreme could be observed if the null hypothesis were true. Small P-values suggest that the null hypothesis may be false.
4. The null hypothesis (no difference) is rejected in favor of the alternate hypothesis (i.e., *B* is more effective than *A*) if the P-value is $\leq \alpha$, the *significance level*. Values for α are small, typically .05 and .1, to reduce the chance of a Type I error.



“A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically non significant and important.”

Which test to use depends ...

- Commonly used in IR papers:
 - Mann-Whitney-Wilcoxon test (Wilcoxon Rank-Sum test)
 - Wilcoxon signed rank test (paired)
- Software packages exist in R, Python, SPSS, etc. that help you test
- A good book to find the right test for a given scenario



User-centered system evaluation

What if we are no longer happy to consider the toy Eiffel tower only?

Lets evaluate real systems with **real users** (=people using the system) **at small scale**.

Instead of "is the system any good?" we are now interested in "can users use the system to retrieve any good results?"



Relevant Factors in Interactive IR (or IIR)

- **Physical, cognitive and affective:** satisfaction with the system, difficulty of use (cognitive load), feelings after usage, etc.
- **Interactions between users and systems:** number of clicks, number of queries issued, query length, etc.
- **Interactions between users and information:** dwell time on a document, terms extracted from a snippet and used in a query, etc.

IIR approaches are diverse

- An **evaluation** measures the quality of a system, interface widget, etc. while an **experiment** compares at least two items (usually a baseline and an experimental system) with each other
- **Lab** (lots of control but artificial), **online** (some control, still artificial) vs. **naturalistic** (little control) studies
- **Longitudinal** studies: require an extended period of time (e.g. *investigate how students interact across 10 weeks with search engine X during their literature survey*)
- **Wizard of Oz** study: participants interact with a system they believe to be automated (in reality it is operated by a human)

Variables

Independent variables: the causes

E.g. investigate how young an old people use an experimental and baseline IR system.

→ age is the independent variable

Dependent variables: the effects

E.g. satisfaction with the search systems

Confounding variables

Affect the independent and dependent variables, but have not been controlled by the experimenter.

E.g. older people are not as familiar with the experimental device as young people.

The experimental design in IIR examines the relationship between 2 or more systems (independent variable) on some set of outcome measures (dependent variables).

Measurements

- “Query logs” or “transaction logs” are usually analyzed
- What can and should be measured depends on the research questions and the setup of the experiment (in a lab or online?)
- Logging clicks is insufficient as user studies have usually few participants (in contrast to Google/Bing with billions of clicks per day)
- Client-side logging is often necessary to track mouse hovers, document dwell time, eye movements (can be done via the Webcam), user activities in other browser tabs/windows, rephrasing of queries, ...

Todos:

**Form a group. Sign up via
Brightspace. Start
discussing your proposal.**

Slack: `in43252019.slack.com`

Email: `in4325-ewi@tudelft.nl`

Today:

Basics of evaluation in IR.

Tomorrow:

Retrieval models.